MINISTERUL EDUCAȚIEI, CULTURII ȘI CERCETĂRII AL REPUBLICII MOLDOVA

Universitatea Tehnică a Moldovei

Facultatea Calculatoare, Informatică și Microelectronică

Departamentul Ingineria Sotfware și Automatică

Admis la susținere

Șef department:

Ion Fiodorov, conf. univ, doctor

_____

"_____"_____2022

Analiza afectării stării emoționale în dependență de contextul scrierii la tastatură

Teză de master

Student: _____ **Tiora Alexandru, IS-211M**

Coordonator: _____ **Grozavu Nistor, conf. univ., doctor**

Consultant: _____ **Catruc Mariana, lect. univ**

Chișinău, 2022

REZUMAT

Primele progrese în colectarea data cu ajutorul tehnologiilor digitale au dus la un volum foarte mare de date. Cea mai mare parte a informației de pe internet este formată din text nestructurat ai semistructurat. Găsirea unor tendințe potrivite pentru analizarea documentelor text din volumul mare de data - este o problemă. Text mining este a proces extragere tendintelor triviale si nontriviale din volumul mare de texte.

Propunerea acestui proiect de master este de a extrage informații relevante, de a găsi tendințe, de a analiza rezultatele din texte scrise de oameni, pentru a înțelege comportamentul acestora. De altfel, este foarte interesant să înțelegem cum influențează contextul informației stamentul emoțional al persoanelor care scriu acest text. Pentru rezultate mai detaliate s-a propus analiza textul scris de copii, deoarece aceștia sunt predispuși să emită emoții mai intense.

ABSTRACT

Fast progress on collecting data using digital technologies brought to very huge volume of data. The most information on the internet consists unstructured and semi-structured text. Finding appropriate trends for analyzing text documents from big volume of data – is a problem. Text mining is a process about extracting interesting and nontrivial patterns from big volume of texts.

The propose of this diploma project is to extract relevant information, finding trends, analyze results from preparing text written by people, for understanding their behavior. For more is very interesting thing to understand how does the context of information affects emotional statement of people who write this text. For more detailed results was proposed to analyze text written by children because they're prone to emit emotions more intense.

Dictionary

NLP	Natural language processing
NLTK	The natural language toolkit
SOM	Self-Organizing Map
VADER	Valence aware dictionary and sentiment reasoner
CBOW	Continous bag of words
SOFM	Self-Organizing Feature Map
ANN	Artificial neural network
t- SNE	t-distributed stochastic neighbor embedding

Contents

Conclusion

During the analysis of the problem and the search for solutions it was assumed that the result would be a model that can group words depending on the burst value of the text sequence. This was partially executed, which shows us the top words of the WordCloud graphs. In 4 out of 5 clusters the word violence predominates.

This can be caused by several reasons. The first reason could be that the word was written several times with both minimum and maximum burst values. In this case it is necessary to perform BERT vectorization which will take into account the written context compared to FastText. Another case may be that the model configuration. Unfortunately at the moment there is no bigger data to apply unsupervised learning with more massive information which could affect the results of the analysis.

After the execution of the work, I gained new knowledge in the field of research and analysis of NLP methods, especially FastText and BERT vectorization. Both methods were compared in the result. Knowledge of new text analysis tools was gained. For example with the help of VADER sentiment analysis it is possible to check what is the status of a sentence or text sequence. In relation to BERT, the result is more significant for this method compared to FastText. Conclusions were also drawn on the importance of pre-processing the text. The initial dataset, having a line count of about 15000 lines, was reduced to 4700 lines because most of the information did not correspond as relevant information.

Bibliography

1. TURNEY,P.D.,PANTEL,P. *From Frequency to Meaning: Vector Space Models of Semantics. Journal of Artificial Intelligence Research.* 2010. 37, 141-188 p. Disponibil: https://doi.org/10.1613/jair.2934
2. MIKOLOV,T.,CHEN, K.CORRADTO.G., DEAN, J. *Efficient Estimation of Word Representations in Vector Space. Proceeding of the International Conference on Learning Representations Workshop*, Arizona, USA, 2013, 1-9 p. Disponibil: https://arxiv.org/pdf/1301.3781.pdf
3. BOJANOWSKIU, P., GRAVE, E., JOULIN, A., MIKOLOV, T. *Enriching word vectors with Subword information. Transactions of the Association for Computational Linguistics*, 2017, 135-146 p. Disponibil: https://doi.org/10.1162/tacl_a_00051
4. MIKOLOV, T., KOPECKY, J., BURGET, L., GLEMBEK, O. CERNOCKY. J. . *Neural network based language models for higly inflective languages*, 2009, ICASSP
5. LIU, Y., OTT, M., GOYAL, N., DU, J., JOSHI, M., CHEN, D., LEVY, O., LEWIS, M., *Robustly Optimized BERT Pretraining Approach*, 2019 Disponibil : https://arxiv.org/abs/1907.11692
6. COLLOBERT, R., WESTON, J., *A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In International Conference on Machine Learning*, 2008 .
7. DUDA, R., HART, P., STORK, D., *Pattern Classification* ,3d ed. California, 2013
8. WITTEN, I., EIBE, F., HALL, M., *Data Mining: Practical Machine Learning Tools and Techniques* , 3d ed., 2011, ISBN 978-0-12-374856-0
9. CHAKRABARTI, S., *Mining the Web: Discovering Knowledge from Hypertext Data*, 2d ed., 2013, ISBN-13 978-1558607545